基于 Sentence2vec 与半监督算法的中文问答提问模式抽取

张金壬，章　韵，王　宇

(南京邮电大学 计算机学院、软件学院、网络空间安全学院，南京 210003)

摘　要：关系抽取是信息抽取中一项重要任务，在处理问答对形式的文本时，除了文本中实体间的关系抽取之外，作为连接问句和答句之间关系的提问模式同样需要抽取。通过有监督的标注算法(条件随机场)与基于模板元组自举的半监督算法的结合在抽取实体间关系时有不错的表现。但传统半监督中发现句式模板的方式难以迁移到提问模式抽取中，针对这种情况通过引入句向量计算文本相似度并选取句式模板，提出一种基于 sentence2vec 技术与半监督算法结合的模型。对于最终实验，采用随机抽样进行验证。实验结果表明，相较于传统的半监督算法，本文的方法得到了更高的准确率和召回率。

关键词：关系抽取；提问模式；条件随机场；自举；句向量

中图分类号：TP391　　doi: 10.3969/j.issn.1001-3695.2018.01.0020

Question pattern extraction based on Sentence2vec and semi-supervised algorithm for Chinese Q&A

Zhang Jinren, Jin Yun, Wang Yu

(*School of Computer Science*, *Nanjing University of Posts & Telecommunications*, *Nanjing* 210003, *China*)

Abstract: Relation extraction is an important task in information extraction. While dealing with the question-answer pairs, in addition to the relations among the entities in the texts, the question pattern as the relation connected questions and the answers also needs to be extracted. The combination of the supervised labeling algorithm (conditional random field) and the semi-supervised algorithm based on a feature template (bootstrapping) has a good performance when extracting relationships between entities. However, the method to find the template in the traditional semi-supervised algorithm was hard to move to the extraction of the question pattern. Therefore, a model based on the combination of sentence2vec technology and semi-supervised algorithm is proposed, which introduce the sentence vector to calculate the text similarity and select the sentence template. Random sampling validation is used to verify the final result. The experimental results show that the method has higher precision and recall values than the traditional semi-supervised algorithm.

Key words: Relation extraction; Question pattern; Conditional random field; Bootstrapping; sentence2vec

0　引言

随着移动互联网的普及与发展，大量结构各异、不同领域的文本不断涌现。为了从这些开放式的文本中抽取非限定类型的关系实例，开放式关系抽取的概念被提出[1]。关系抽取作为信息抽取中的重要步骤，最终目的是建立文本中实体或事件间语义逻辑上的关联，并形成结构化表示[2, 3]。问答对是其中一种特殊形式的文本。作为人工客服记录、社区论坛页面文档、智能搜索引擎等众多类型文本的载体，问答对包含着丰富的信息与知识，面向问答对的开放式关系抽取正成为业界研究的热点。

问答对是一种上下文具有逻辑关联性的文本，孤立的问句或者答案在内容和逻辑上是缺失的。而连接问句与答案的关系就是提问模式。针对问答对形式的文本，关系抽取包括提问模式抽取（即问题与答案之间的关系）及内容关系抽取（即实体与实体间的关系）[4]。目前，针对实体间关系抽取的研究较为成熟，但将部分算法迁移到提问模式抽取时，算法的性能会下降。中文问答对当中的提问模式抽取挑战在于，除去一些规整的特殊字符，提问模式的表述方式会有很多，甚至相互文本之间的间隔很远。另外中文当中的一词多义情况也是需要克服的问题。完善面向问答对的关系抽取技术，对理解用户语义和提问意图、构建智能问答系统、促进建立知识库等方面有着重要意义[5]。

目前提问模式抽取的方法主要分为基于知识工程和基于

收稿日期：2018-01-10；**修回日期：**2018-03-08

作者简介：张金壬（1991-），男，江苏扬州人，硕士研究生，主要研究方向为自然语言处理（zhangjinren@aliyun.com）；章韵，男，江苏南京人，教授，硕导，博士研究生，主要研究方向为计算机通信、无线传感及云计算；王宇（1992-），男，江苏南京人，博士研究生，主要研究方向为自然语言处理.

机器学习两种方法。在面向开放式问答对的抽取过程中，需要充分考虑文本的冗余性以及抽取方法的轻量化，因此，基于规则模板和模式学习等知识工程方法[6]由于本身移植性和覆盖率等问题无法很好的解决开放式提问模式抽取。

基于机器学习的方法可以分为三种：有监督算法、半监督算法和无监督算法。在有监督学习中，一些传统算法能够准确抽取出关系元组，如基于图模型的算法 CRF[7]，但算法本身过于依赖特征模板构建以及大量人工标注量，当数据量增大时，算法无法有效覆盖数据。半监督算法能够有效提升召回率，经典的如 DIPRE[8]，该方法的关键在于如何发现包含正确元组的句子模式，传统做法，出现“语义漂移”[9]的现象。无监督算法的思路是对可能包含关系词的文本进行聚类，将聚类结果中的高频词作为关系类型[10]。无监督学习的问题在于对于关系聚类的结果无法定义，而且对于低频的关系表述难以抽取，因此无监督算法一般不独立解决抽取问题。无监督算法中基于深度学习的文本聚类思想与其他机器学习算法结合是目前主流的方法，其中基于稀疏表示的分类（SRC）方法在模式识别和机器学习方面取得了许多成功[11]。sentence2vec[12]就是一种轻量级基于深度学习的无监督稀疏表示算法，本文将利用其在计算文本相似度方面的优异性能用于改进半监督算法。

为了提升提问模式抽取的性能，本文提出一种基于 sentence2vec 的半监督算法模型。在提问模式的关系元组中，提问模式的两端可能是实体，也可能是多个实体与关系构成的事件。为了便于描述，结合 TAC 对事件的定义，本文将问题中除了提问模式的字符串序列（即对提问内容的描述部分）称为提问事件 E_1，答案字符序列称为对 E_1 的答案事件 E_2，提问模式为两者之间的关系 R，最终将一条问答对生成形如(E_1，R，E_2)的三元组。本文的方法首先通过有监督算法得到提问模式的种子集，再通过半监督算法用以扩充抽取元组量，对传统半监督算法中通过匹配实体发现句式的方法提出了改进，利用 sentence2vec 技术与半监督算法结合，与传统方法对比实验结果表明，本文的算法抽取的正确提问模式数量明显增加，覆盖率也有了明显的提升，有效的提升了半监督算法的性能。

1　相关工作

早期的关系抽取主要依赖于预定义的关系类型，MUC-7 会议上首先提出了 Location_of、Employee_of、Product_of 三种面向商业活动内容的关系[13]。之后的 ACE 会议又将关系种类定义为包括机构关系、整体部分关系、人-社会关系在内的七大类关系。抽取方法也从模式匹配转向机器学习。这些研究在某些特定领域取得了不错的效果[13]。在面对 Web 页面海量数据集时，事先定义好关系类型的抽取方法就很难胜任了。随着 OpenIE[14]概念的提出与兴起，机器学习方法尤其是深度学习方法的开始成为研究的热点。

在英文领域中，开放式实体关系抽取的相关研究与技术已经有不少成果， Brin 等利用半监督学习的方法，设计出 DIPRE（Dual Iterative Pattern Relation Expansion）系统。该系统通过对少量种子模板的不断迭代，实现了自动抽取 Web 页面上的实体信息和实体关系。Banko 等人首次提出了 OpenIE（OIE）的概念，他设计的 TextRunner 系统采用启发式规则进行自动标注得到种子模板，并对种子模板学习生成分类器，进而进行抽取[15]。Wu 等人提出的 WOE 又借助维基百科的条目属性等信息进行标注，提高了标注的质量[16]。

中文领域中的关系抽取目前也取得了很多成果。吴友政等提出一种基于无监督算法的提问模式抽取方法，结合模板匹配，在面向开放式文本抽取时，取得了不错的效果[17]。刘安安等人设计了面向句子级的开放式关系抽取系统 TMS，该系统通过启发式的模板对其进行筛选，在句子级的关系抽取比传统的方式性能有所提升[18]。王明印等人提出了 SCOERE 的半监督开放式关系抽取方法，对的句子进行二元实体关系进行标注，并通过自学习的方式提高了监督学习方法的性能[19]。

2　提问模式抽取模型

2.1　模型描述

图 1 是本文提出的开放式提问模式抽取模型，核心部分为基于 sentence2vec 技术的自动标注与半监督算法 Bootstrapping 结合的闭环模块。原始语料通过数据预处理部分得到文本特征并进行词向量（word embedding）的构建。文本特征与人工标注用于生成有监督学习模型 CRF，模型产生提问模式的种子集。种子集与词向量作为自动标注的输入用于发现同义句式，结合半监督算法不断迭代抽取新的提问模式，选取其中高置信度的元组输出。接下来本章将详细介绍每个模块的原理及实现过程。

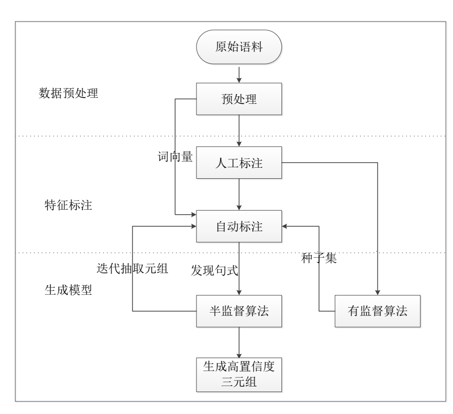


图 1　提问模式抽取模型

2.2　基于监督算法 CRF 的提问模式抽取

基于马尔可夫随机场的 CRF 是目前主流的标注算法，在诸如图像检测与中文分词领域都取得了不错的效果[20]。其本质是在给定的随机变量 X 时，随机变量 Y 的马尔可夫随机场。如在线性链上，X 表示需要标注的观测序列(X_1,X_2 ,X_3……,X_n)，Y 则是输出的状态序列(Y1, Y2, Y3……, Yn)，在给定输入数据

X 时，通过条件随机场的条件转移概率模型 P (y|x)得到输出序列 Y。条件转移概率模型如下：

$$p(y|x)=\frac{1}{Z(x)}\exp\left(\sum_{i,k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i,l}\mu_l s_l(y_i,x,i)\right) \quad (1)$$

$$Z(x)=\sum_y \exp\left\{\sum_{i,k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i,l}\mu_l s_l(y_i,x,i)\right\} \quad (2)$$

CRF 同时支持多元特征扩展，扩展关键在于两点：上下文观测窗口和特征选择。在观测窗口方面，作为典型的 n-gram 模型，n 为观测窗口。n 越大，效果越好，但同时计算复杂度也越高。一般来说，当 n>4 时，提升的效果趋于平缓。在本文第四章的实验中，会详细介绍窗口大小调优过程。

在特征选取方面本文选取了以下几个特征：词语特征、词性特征、实体类型、依存路径和依存关系类型。实体类型是指实体所属类型，分别是领域内实体类型、机构名实体类型、时间类型、地点类型及其他。词语的依存路径特征是由两项特征表示：1.词语是否为其他词语的依存词，2.当前词语所在依存树的层数。

人工标注除了传统的 B、I、E、O 标注类型之外。对于提问模式缺省的文本，本文在句尾的标点上标注为 L（lack）以区分陈述句。

2.3 基于半监督算法 Bootstrapping 的提问模式抽取

有监督算法 CRF 可以取得不错的准确率，但召回率较低。半监督学习方法 Bootstrapping 在传统关系抽取中表现出不错的召回率，以少量已标注种子集为输入，利用模式元组二元性，不断迭代发现包含新关系句子模式，再从句式中得到可信的关系元组，在本文中即为提问模式。

Bootstrapping 的模式元组二元性是指，以实体为窗口的三元组上下文会形成一个句子模式，新的关系描述可能来自相同的句子模式；相反的，一个可信的三元组一定会出现在不同的句式中，因此理论上当有一个理想的生成句式的方法时，就可以借助少量的种子元组不断发现不同的句式，继而抽取更多的元组。Bootstrapping 算法如下：

算法 1：Bootstrapping

输入：少量带标注的语料 T_L，大量未标注语料 U

输出：大量可信的带标注 T_L

```
1) Foreach T_L
2)     R←Pretreatment(T_L)
3) S←FindSentence(R, U)
4) P←Generate(S)
5) R’ ←M_D(P)
6) End foreach
```

其中，2)是从标注语料 T_l 中提取提问模式元组 R。3)是从未标注语料中找出包含元组的上下文 S。4)是通过 S 构造句子模式，即在上下文中选定包含元组信息的内容，这也是算法核心部分。传统算法通过语法和领域知识制定规则确定内容，生成一个形如 3.1 的七元组：

$$(left,tag_1,mid_{12},tag_2,mid_{23},tag_3,right) \quad (3.1)$$

在这里 tag_i（i=1,2,3）是指实体及其之间的关系描述，在本文即为事件与提问模式。而 left 和 right 则为上下文，mid 为实体间的文本。5)是从未标注语料 U 中提取符合模式 P 的高置信度的元组 R’。本文将句式中实体及停用词之外的部分标注为提问模式，选取高置信度的元组，置信度的计算公式为 $\mathrm{conf}(K)=1-\prod_{i=0}^{|p|}\frac{1-prob(p_i)}{E_{sel}}$，其中 E_{sel} 表示模式 p_i 抽取的关系元组的数量，$p=\{p_i\}$ 表示抽取到的提问模式 K 的句子模式。简而言之，若同一个提问模式可以从很多个句子模式中抽取出来，便被认为是一个可信的提问模式。接下来将介绍本文对构造句式部分的改进。

2.4 基于 sentence2vec 的语义相似度计算

提问模式的两端并不以显性的实体作为固定上下文窗口的标识，因此传统的 Bootstrapping 利用实体作为模板窗口标识的方法不能有效的迁移到提问模式抽取的问题上。构造句式目的在于同义句中对不同提问模式的表述进行抽取，本文提出一种基于 sentence2vec 的语义相似度计算，用于得到同义或近义的句子模式。

传统的计算句子相似度一般是将句子中的词语映射到 one-hot 形式的向量中，通过计算句子间的编辑距离（包括欧氏距离，余弦相似度等）得到结果，这类基于字符或词语的算法无法表示上下文的关联，因此无法真正得到语义相似的句子。

而基于深度学习得到的语言模型能够较好的表示上下文词语的联系。其中，一种轻量级的 sentence2vec 在对句子级文本进行高维向量映射表现出了不错性能。在该方法中，每个句子的语义向量 c_s 的语义概率模型为

$$p(s|c_s)=\prod_{w\in s}p(w|c_s)=\prod_{w\in s}\left[\alpha p(w)+(1-\alpha)\frac{exp(v_w,c_s)}{Z}\right] \quad (1)$$

其中：w 表示当前词语，c_s 表示当前句向量表示，α 为自定义常量，v_w 表示当前词的词向量表示，$Z=\sum_{w\in s}exp(v_w,c_s)$ 表示之前词向量与句向量的内积和，也是一个常量。对（3）式取对数，得到 s 的最大似然估计 $f_w(c_s)=\log\left[\alpha p(w)+(1-\alpha)\frac{exp(v_w,c_s)}{Z}\right]$，按泰勒展开式得出：

c_s 的极大似然估计正比于所有词向量的加权求和：

$$\arg\max\sum_{w\in s}f_w(c_s)\propto\sum_{w\in s}\frac{a}{p(w)+a}v_w, where\, a=\frac{1-\alpha}{\alpha Z} \quad (2)$$

词向量 v_w 的权值称为 “逆光滑频率”，权值中包含两个参数 $p(w)$ 和 a，在该方法中，还引入公共句子向量 c_o 的概念，来表述所有句子共有的冗余语义，因此最终的句子向量 v_s 中需要减去 c_o 的部分。Sentence2vec 算法如下：

算法 2： sentence2vec

输入：词向量 $\{v_w: w\in V\}$，问句集 S，参数 a 以及每个词出现的概

率估计 $\{p(w): w \in V\}$

输出：相似句子向量 $\{v_s : s \in S\}$

```
7) for all sentence s in S do
```

$$c_s = \frac{1}{|s|}\sum_{w \in s}\frac{a}{p(w)+a}v_w$$

```
8) end for
9) Compute the first principal component u of {v_s : s∈S}
10) for all sentence s in S do
```

$$v_s = c_s - c_o c_o{}^t c_s$$

```
13)              end for
```

其中，8）表示生成当前语义向量，为了求解参数 a，对比 word2vec 的 CBOW 模式中的语言模型：

$$\Pr[w_t \mid w_{t-1},\ldots,w_{t-5}] \propto exp(\bar{v}_t, v_w), where \bar{v}_t = \frac{1}{5}\sum_{i=1}^{5} v_{w_{t-i}} \qquad (3)$$

$w_{t-1},\ldots,w_{t-5}$ 表示当前语境（设句子长度为 5），w_t 当前语境的下一个单词，整个语言模型正比于下一个词向量 v_w 和语境中平均向量 $\bar{v}_t$ 的内积。在这里，每个词向量的极大似然估计可以表示为：

$$g(v_w) = \sum_{(w,c)\in D} \log\sigma(\bar{v}_t \cdot v_w) + \sum_{(w,c)\in D`} \log\sigma(-\bar{v}_t \cdot v_w) \qquad (4)$$

式（4）的第二项为负采样项，$\sigma(x)$ 为 sigmoid 函数，$D`$ 为新出现非 v_w 集合，求解 $f(v_w)$ 的参数 $\bar{v}_t$ 时，需要通过随机下降的方式确定梯度，句子中每个词向量的参数是一个贝努力分布 $q(w_{t-i}) \triangleq \left\{1, \sqrt{\frac{10^{-5}}{p(w_{t-i})}}\right\}$，则新出现的词 v_w 的期望可以表示为其前文词向量的加权求和：

$$E(g(v_w)) = \alpha\begin{pmatrix} q(w_{t-5})v_{w_{t-5}} + q(w_{t-4})v_{w_{t-4}} + q(w_{t-3})v_{w_{t-3}} + \\ q(w_{t-2})v_{w_{t-2}} + q(w_{t-1})v_{w_{t-1}} \end{pmatrix} (5)$$

式(5)在形式上可以类比同样采用随机梯度下降的式(1)。而在本次实验的数据集上随机抽样了 500 个词，当 a =0.001 时，$q(w_{t-i})$ 与式（3）中的权值几乎相同，在句子相似度表现最好。

在得到句向量之后，通过皮尔森相关系数计算句向量之间的相似度。对相似度降序排序，计算相邻结果相似度的差值，当差值超过阈值 τ=0.2 时，认为语句不相似，将之前的语句输出作为相似语句，作为生成新元组的模式。

3　实验及性能分析

3.1　实验数据及性能指标

本章将针对上一章介绍的模块进行调优和对比实验。本文实验数据通过网络爬虫从 Web 页面获取中文问答对（FAQs），主要有三个来源：a）电信 10000 号 FAQs，83121 条；b）百度知道相关类目问答，37400 条；c）新浪爱问知识人相关类目问答，43020 条。在实验前清洗了一些噪声数据：不包含相关领域实体的问句；含有特殊字符的问答对;统计发现，一般问句不会超过 30 个词以上，主要分布在 15 个词左右，因此字数超过 30 的句子也舍弃，最终整个数据包含 129937 条问答对。

本文的实验包括以下四个：a)测评 CRF 抽取提问模式的性能；b)对比 sentence2vec 与传统方法性能；c)测评本文提出的结合 sentence2vec 的半监督算法抽取标注提问模式效果性能；d)比较本文方法与基于规则匹配句式的传统 Bootstrapping 算法之间的性能。本文的测评标准包括：准确率 $P = \frac{正确识别的元组个数}{识别的元组个数}$，召回率 $R = \frac{正确识别的元组个数}{数据集中元组个数}$，$F值 = \frac{2PR}{P+R}$。

3.2　CRF 标注性能分析

本文采用的 CRF 通过开源框架 sklearn-crfsuite 实现。词语本身和词性特征作为原始的两项特征。为了验证其他几项特征选取对 CRF 标注的影响，本文设定两种情况来验证加入后的测评指标：窗口大小对标注性能的影响；加入多元特征对性能的影响。

样本随机选取了 500 条语料，窗口大小对性能的影响如表 1 所示。其中，“1W”是指观测窗口大小为 1，观测一维特征。“1W+2”是指观测窗口大小为 1，同时观测两个词的特征，以此类推。结果显示窗口大小为 2 并同时观测两词特征和三词特征的情况表现较好，准确率分别达到了 66.31%和 66.62%，接下来以这两项继续添加多元特征来观测性能。

表 1　不同窗口大小及观测数量对 CRF 标注结果的影响/%

窗口	1W	2W	3W	1W+2	1W+3	2W+2	2W+3	3W+2	3W+3
准确率	62.64	63.39	63.51	66.27	65.37	66.31	65.62	64.31	64.31
召回率	35.74	37.84	37.33	38.72	38.74	38.70	39.02	38.73	38.43
F 值	46.77	47.39	47.02	48.89	48.64	48.88	48.93	48.34	48.11

这里用 F 代表候选的特征。实体类型(F1)、当前词所在依存树层数(F2)、是否是其他词的依存词(F3)、当前词的句法分析路径(F4)以及依存语法类型（F5）。结果如表 2、3 所示，同时选取实体类型，依存路径，依存类型作为特征在标注关系时取得了最高的性能。其中，F2 和 F3 是作为整体表征当前词在依存树中的位置的，而依存关系对于性能提升明显，在分析数据后发现，有些提问模式在句中表现形式并不是连续的，需要词语之间相互依存关系作为特征。而实体类型作为对词性特征的补充描述，也起到了一定提升性能的作用。

依存类型则是表现语法对于提问模式的作用。但句法分析路径并没有对整个标注性能有太大的提升。考虑模型训练时间，本文最终舍弃这个特征。根据以上实验，最终的窗口大小为 2，特征为词语字符、词性、实体类型、依存路径和依存关系类型。

表 2　窗口 2w+2 不同特征模板对 CRF 性能的影响

模板	F1	F123	F1234	F12345	F1235
准确率	72.60%	81.62%	81.8%	82.78%	82.52%
召回率	38.84%	41.33%	41.72%	41.74%	41.70%
F 值	47.39%	47.02%	48.89%	48.64%	48.88%

表 3 窗口 3w+3 不同特征模板对 CRF 性能的影响

模板	F1	F123	F1234	F12345	F1235
准确率	68.6%	63.51%	66.27%	65.37%	66.31%
召回率	37.84%	37.33%	38.72%	38.74%	38.70%
F 值	47.39%	47.02%	48.89%	48.64%	48.88%

将模型在整个数据集上进行实验，本文采取了十折交叉验证的方案，并对结果随机抽样 1000 条进行人工验证。结果如图 2 所示，准确率为 83.7%，但召回率仅为 31.8%，基于有监督的算法抽取的提问模式基本准确，但覆盖率不高，下面将会实验半监督算法对性能的提升效果。

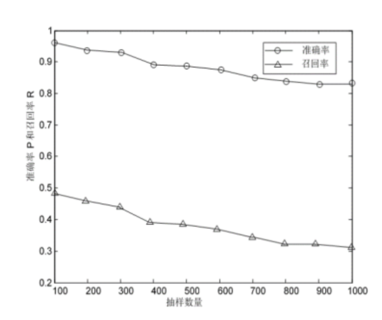


图 2 有监督算法提问模式抽取的性能

3.3 语义相似度性能分析

在分析半监督算法性能之前，本文将对比对 sentence2vec 的性能。本文的深度学习模型都基于 TensorFlow 框架实现，数据为原始语料中随机抽样 2000 对问句，这里主要关注的是除实体外的提问模式语义是否相似。本文采用了一些其他方法进行对比，包括基于 TextRank 相似度、基于 LDA 相似度和基于 word2vec 线性相加相似度，结果与人工判别相似结果对比。

从表 4 可以看出，基于深度学习语言模型 sentence2vec 的语义相似度计算在准确率上的表现要明显优于基于 TextRank 的相似度计算和基于 word2vec 线性相加的相似度，也比基于 LDA 相似度的准确率略高。

表 4 基于 sentence2vec 语义相似度与其他算法性能对比

方法	正确匹配的问句	准确率(%)
基于 TextRank 相似度	654	32.7
基于 LDA 相似度	1526	76.3
基于 word2vec 线性相加相似度	1242	62.1
基于 sentence2vec 相似度	1747	87.4

3.4 自动标注元组抽取性能分析

通过自动标注得到了更丰富的标注预料，对最终结果随机采样 1000 条进行人工验证。图 3 为标注的性能展示，准确率为 80.3%，召回率为 71.6%，F1 值为 75.7%。可以看到最终得到在加入了自动标注的结果后，召回率有了明显的提升，虽然在最终结果的准确率上略有下降，但正确的关系实例数量显著提高，F1 值的明显提升充分证明半监督算法对提高抽取覆盖率的想法是可行的。

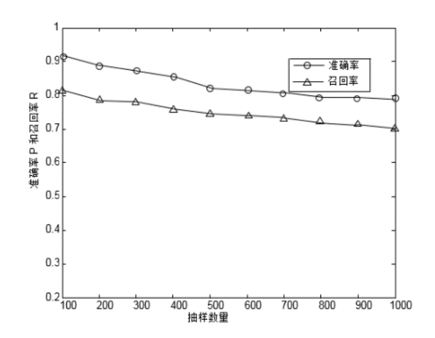


图 3 结合 Bootstrapping 自动标注对性能的影响

3.5 改进的 Bootstrapping 算法性能对比分析

受到 Zhao 等工作的启发，本文设计对比实验，将传统 Bootstrapping 算法迁移到提问模式抽取的问题上，具体步骤如算法 1 所示[21]。在关键步骤 4)中，以七元组形式将实体上下文作为句式模板进行匹配，对比本文提出的改进算法。实验数据仍采用的 1000 份抽样样本人工比对，准确率和召回率对比如图 4、5 所示。

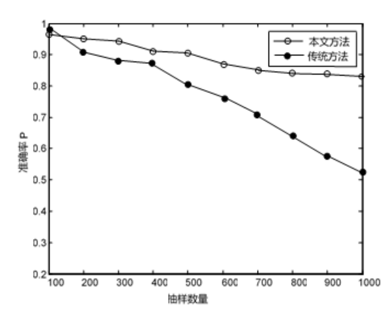


图 4 改进后的 Bootstrapping 于传统方法准确率对比

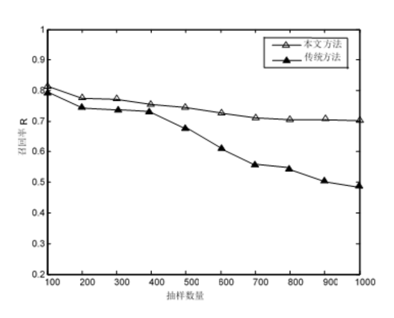


图 5 改进后的 Bootstrapping 于传统方法召回率对比

可以看出，传统的 Bootstrapping 算法迁移效果并不理想，在少量实体相同的文本中虽然还可以保证一定的准确率召回率，但随着样本量变大，抽取的准确关系实例数量迅速变少，出现了“语义漂移”的现象。本文提出的算法针对同义句式匹配作出了改进，保证了在抽取过程中，发现句式的稳定性，使得性能在数据量增长的同时仍能保持较好的准确率和召回率，实验结果证明，基于 sentence2vec 的半监督算法，在抽样样本上的准确率达到了 83.5%，召回率达到了 70.3%。相较于传统的方法，当数据量上升时，性能得到了有效的提升。并且由于不再依赖文本当中的实体，因此，本文的方法也具有

较好的可移植性和鲁棒性。

4　结束语

本文提出了一种面向中文问答对的提问模式抽取的方法，将提问模式看做两个事件之间的关系进行抽取。采用监督学习的算法得到提问模式的种子集，再通过 bootstrapping 算法，利用模式元组的二元性得到更多的提问模式。针对问答对中不存在显性实体对的情况，本文提出了一种基于 sentence2vec 算法的文本相似度计算方法，用于得到可信的句子模式，从而有效的提高了提问模式抽取的准确性与泛化性，在 Web 网页中随机抓取的电信领域文本实验结果也证明了本文方法的有效性，此方法也可以推广到其他领域的问答对的抽取问题中。接下来会在更广泛的 Web 页面文本上进行实验，寻找更多表征提问模式的特征，设计更好的自动标注算法，以期覆盖更丰富类型的文本。

参考文献：

[1] Banko M, Cafarella M J, Soderland S, *et al.* Open information extraction from the Web [C]// Proc of IJCAI. 2007, 7: 2670-2676.

[2] 黄勋，游宏梁，于洋. 关系抽取技术研究综述 [J]. 现代图书情报技术, 2013 (11): 30-39.

[3] Bouma G, Fahmi I, Mur J. Relation extraction for open and closed domain question answering [C]// Theory & Applications of Natural Language Processing. 2011: 171-197.

[4] Tushar K, Ashish S, Peter C. Answering complex questions using open information extraction [J/OL]. (2017-04-19) . https://arxiv. org/pdf/1704. 05572. pdf

[5] 余正涛，毛存礼，邓锦辉，等. 基于模式学习的中文问答系统答案抽取方法 [J]. 吉林大学学报：工学版, 2008, 38 (1): 142-147.

[6] Brin S. Extracting Patterns and Relations from the World Wide Web [C]// Proc of International Workshop on the World Wide Web and Databases. Berlin: Springer, 1998: 172-183.

[7] Curran J R, Murphy T, Scholz B. Minimising semantic drift with Mutual Exclusion Bootstrapping [EB/OL]. 2007: 172-180. http://www. it. usyd. edu. au/~james/pubs/pdf/pacling07boot. pdf.

[8] Hasegawa T, Sekine S, Grishman R. Discovering relations among named entities from large corpora [C]// Proc of Meeting on Association for Computational Linguistics. [S. l.] : Association for Computational Linguistics, 2004: 415.

[9] Adi Y, Kermany E, Belinkov Y, *et al.* Fine-grained analysis of sentence embeddings using auxiliary prediction tasks [J]. arXiv: 1608. 04207 2016.

[10] Wang Z, Teng S, Liu G, *et al.* Hierarchical sparse representation with deep dictionary for multi-modal classification [J]. Neurocomputing, 2017, 253 (C): 65-69.

[11] Kiros R, Zhu Y, Salakhutdinov R, *et al.* Skip-thought vectors [C]// Proc of International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2015: 3294-3302.

[12] Chinchor N, Marsh E. Muc-7 information extraction task definition [C]// Proc of the 7th Message Understanding Conference. 1998: 359-367.

[13] Doddington G R, Mitchell A, Przybocki M A, *et al.* The automatic content extraction (ACE) program-tasks, data, and evaluation [C]. LREC. 2004, 2: 1.

[14] Yates, Alexander, Cafarella, *et al.* TextRunner: open information extraction on the Web [J]. Journal of the Ceramic Society of Japan, 2007, 96 (3): 1127-1130.

[15] Wu F, Weld D S. Autonomously semantifying wikipedia [C]// Proc of the 16th ACM Conference on Conference on Information and Knowledge Management. New York: ACM Press, 2007: 41-50.

[16] 吴友政，赵军，徐波. 基于无监督学习的问答模式抽取技术 [J]. 中文信息学报, 2007, 21 (2): 69-76.

[17] 刘安安. 开放式中文实体关系抽取研究 [D]. 哈尔滨：哈尔滨工业大学, 2013.

[18] 王明印. 开放式中文实体关系抽取研究 [D]. 北京：北京邮电大学, 2015.

[19] Zhao Z S, Feng X, Wei F, *et al.* Learning representative features for robot topological localization [J]. International Journal of Advanced Robotic Systems, 2013, 10: 1.

[20] Wang Z, Zhao Z, Weng S, *et al.* Incremental multiple instance outlier detection [J]. Neural Computing & Applications, 2015, 26 (4): 957-968.